nature biotechnology

Large-scale genotyping of complex DNA

Giulia C Kennedy[1], Hajime Matsuzaki[1], Shoulian Dong[1], Wei-min Liu[1], Jing Huang[1], Guoying Liu[1], Xing Su[1,4], Manqiu Cao[1], Wenwei Chen[1], Jane Zhang[1], Weiwei Liu[1], Geoffrey Yang[1], Xiaojun Di[1], Thomas Ryder[1], Zhijun He[1], Urvashi Surti[2], Michael S Phillips[3], Michael T Boyce-Jacino[3], Stephen PA Fodor[1] & Keith W Jones[1]

Genetic studies aimed at understanding the molecular basis of complex human phenotypes require the genotyping of many thousands of single-nucleotide polymorphisms (SNPs) across large numbers of individuals[1]. Public efforts have so far identified over two million common human SNPs[2]; however, the scoring of these SNPs is labor-intensive and requires a substantial amount of automation. Here we describe a simple but effective approach, termed whole-genome sampling analysis (WGSA), for genotyping thousands of SNPs simultaneously in a complex DNA sample without locus-specific primers or automation. Our method amplifies highly reproducible fractions of the genome across multiple DNA samples and calls genotypes at >99% accuracy. We rapidly genotyped 14,548 SNPs in three different human populations and identified a subset of them with significant allele frequency differences between groups. We also determined the ancestral allele for 8,386 SNPs by genotyping chimpanzee and gorilla DNA. WGSA is highly scaleable and enables the creation of ultrahigh density SNP maps for use in genetic studies.

We set out to overcome two important bottlenecks in current genotyping technology: the requirement for locus-specific SNP amplification and locus-specific allele discrimination[3,4]. We devised a generic sample preparation method that uses a single oligonucleotide primer for amplification, coupled to allele discrimination on synthetic DNA microarrays. Arrays access large quantities of genetic information by relying on specific hybridization of the nucleic acids in a sample to complementary sequences on the array[5]. Although arrays with >500,000 probe sequences can be synthesized, a principal challenge is to present genomic DNA to the array so as to derive accurate allelic information about the sample. As the number of unique genomic base pairs (complexity) in the target increases, opportunities for cross-hybridization and nonspecific signals increase. It is therefore preferable to present subsets (fractions) of the genome to the array to derive meaningful and specific signals.

There are several well-known methods for genomic fractionation[6–10]. A common theme is the use of restriction enzyme digestion followed by adaptor ligation and amplification with a common primer. The methods differ in the complexity reduction step; for example, representational difference analysis uses PCR properties to selectively amplify fragments up to 1 kb[6,7]. The amplified fragment length polymorphism method uses specific primers to amplify genomic subsets[8]. Genomic subfractions are also prepared by size selection on gels[2,9,10]. The SNP Consortium (TSC) used this approach to identify more than one million SNPs[2,9]. All these methods made advances in the preparation of genomic subfractions, yet they were limited in their ability to discriminate alleles, that is, to genotype, on a large scale.

We required our genotyping approach to satisfy three criteria. First, it should leverage the large numbers of SNPs deposited in public databases by TSC. Second, owing to the large numbers of SNPs being interrogated, it must avoid SNP-specific primers. Finally, it must be highly reproducible and accurate.

To take advantage of the large numbers of SNPs already discovered, we focused on fractionation schemes used by TSC for SNP discovery, namely *Eco*RI, *Bgl*II and *Xba*I restriction enzyme digestion, followed by size selection in the 400- to 800-base-pair (bp) range. However, rather than using gels for size selection, we optimized our amplification conditions to selectively amplify fragments of that size. The biochemical fractionation method we used, 'fragment selection by PCR' or FSP, is shown in **Fig. 1** and described in detail in Methods.

Targets generated by FSP were labeled and then hybridized to arrays. Each *Eco*RI, *Bgl*II and *Xba*I fraction represents approximately 4×10^7 bp of genomic DNA. An image of a representative array hybridized with one fraction showed robust signal intensities (**Fig. 2a**). In contrast, hybridization with an identical mass of total human genomic DNA (3.2×10^9 bp) results in much lower signal intensities (**Fig. 2b**). A close-up view of an SNP 'block' hybridized with DNA from three different individuals representing all three genotypes is shown in **Fig. 2c**. SNPs are genotyped by allele-specific hybridization using an algorithm developed for high-complexity samples[11]. We prepared a target from 108 individuals to train the algorithm. Clusters corresponding to the three possible genotypes were observed in the training set (see **Supplementary Fig. 1** online) and a set of conservative heuristics developed to identify 14,548 high-quality SNPs (Methods). Reproducibility and accuracy across 38 samples was determined (Methods), with an average call rate (number of SNPs identified

[1]Affymetrix, 3380 Central Expressway, Santa Clara, California 95051, USA. [2]Departments of Pathology and Genetics, Magee Women's Hospital, 300 Halket St., Pittsburgh, Pennsylvania 15213-3180, USA. [3]Orchid BioSciences, Inc., 303 College Road East, Princeton, New Jersey 08540, USA. [4]Present address: Biotechnology Research Group, Intel Corporation, 3065 Bowers Ave., Santa Clara, California 95054, USA (X.S.). Correspondence should be addressed to G.C.K. (giulia_kennedy@affymetrix.com).

Published online 7 September 2003; doi:10.1038/nbt869

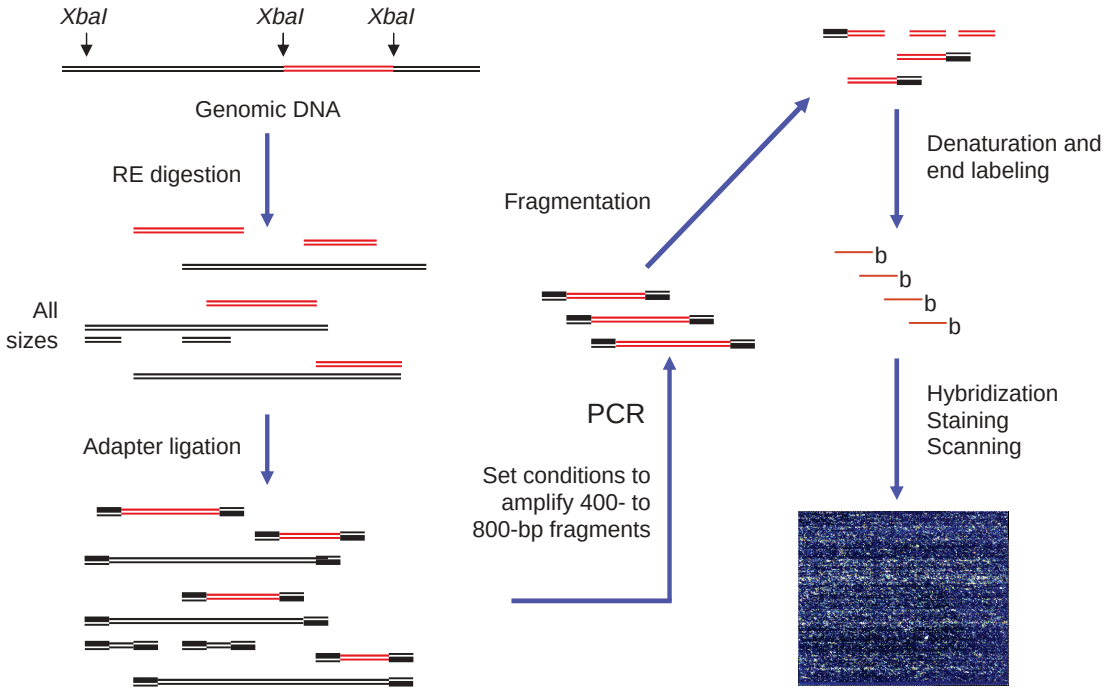

Figure 1 Fragment selection by PCR (FSP). Digestion of genomic DNA with a restriction enzyme (*e.g.*, *Xba*I), results in fragments of various sizes (black), including fragments 400- to 800-bp long (red). Adaptors are ligated to all size fragments, but fragments in the 400- to 800-bp size range are preferentially amplified. The amplified target is fragmented and labeled and hybridized to synthetic DNA microarrays.

divided by the total number examined) of 95.8% and overall concordance rate of 99.1% with an independent genotyping method (**Supplementary Table 1** online).

We examined three genomic fractions, each approximately 43 Mb, hybridized to separate arrays. We then determined the effect of complexity on call rate and concordance by generating a series of increasingly complex target samples from 43 to 425 Mb. We hybridized these to arrays containing SNPs from the *Xba*I fraction and determined call rate and concordancy at three target DNA amounts (**Fig. 3a**,**b**). As complexity increases, both call rate and concordancy decline; however, the effect is more pronounced at smaller target amounts. Thus, it is possible to genotype extremely complex samples (>300 Mb) at >99% accuracy by increasing target DNA amounts.

We used WGSA to determine SNP allele frequencies in DNA from 60 unrelated individuals comprising three human populations: African-American, Caucasian and Asian. A majority of SNPs were polymorphic in all three populations. This is consistent with expectations, as the training set consisted of an ethnically diverse panel of individuals including these three populations as well as two others (Native American and Mexican-American). In this analysis, there were 343, 535 and 1,219 markers in the African-American, Caucasian and Asian samples, respectively, that were monomorphic (that is, had zero heterozygosity).

The distribution of marker heterozygosity in the three populations is shown in **Supplementary Figure 2** online. The mean heterozygosity of the markers is 0.348, 0.354 and 0.322 in the African-American, Caucasian and Asian samples, respectively, indicating that the majority of SNPs are informative in the populations studied here.

We calculated the F_{ST} statistic[12], an estimate of the geographic structure between two populations, for each SNP. F_{ST} values vary from 0 to 1; as allele frequency differences between populations increase, so do F_{ST} values. Mean F_{ST} values are 0.061, 0.094 and 0.065 for African-American versus Caucasian, African-American versus Asian and Caucasian versus Asian comparisons, respectively. Thus the majority of markers show small interpopulation frequency differences (**Supplementary Table 2**). These values are consistent with F_{ST} distributions previously reported for a smaller number of loci in a different set of samples[13]. The comparison between African-American and Asian allele frequencies resulted in overall higher F_{ST} values relative to the other two comparisons. Our results show that whereas most SNPs demonstrate modest allele frequency differences among the three populations, there is a subset of SNPs whose allele frequencies differ substantially in one population versus the other two. These 'ancestry-informative markers' (AIMs), can be used to map complex diseases using admixture-generated linkage disequilibrium (MALD)[14–17]. There are 343, 788 and 374 SNPs with F_{ST} values >0.4 in the African-American versus Caucasian, African-American versus Asian and Caucasian versus Asian comparisons, respectively (**Supplementary Table 2**).

SNPs are sequence changes that arose once during evolution. To determine which allele represents the ancestral state, we genotyped chimpanzee and gorilla DNA samples. Chimpanzee and gorilla DNA differ from human by 1.5% and 2.1%, respectively[18]. Synthetic arrays have been used previously to genotype chimpanzees and gorillas on human SNPs[19,20]. We called chimpanzee and gorilla genotypes on 77.1% and 71.8%, respectively, of the 14,548 human SNPs in this study (data not shown). Almost all markers are called homozygous in both great apes; 97.8% in chimpanzee and 97.7% in gorilla (data not shown), consistent with the recent evolutionary history of SNPs. We assigned ancestral alleles only to SNPs that were homozygous in both chimpanzee and gorilla, and that gave the same genotype in both species. A total of 8,386 SNPs were assigned.

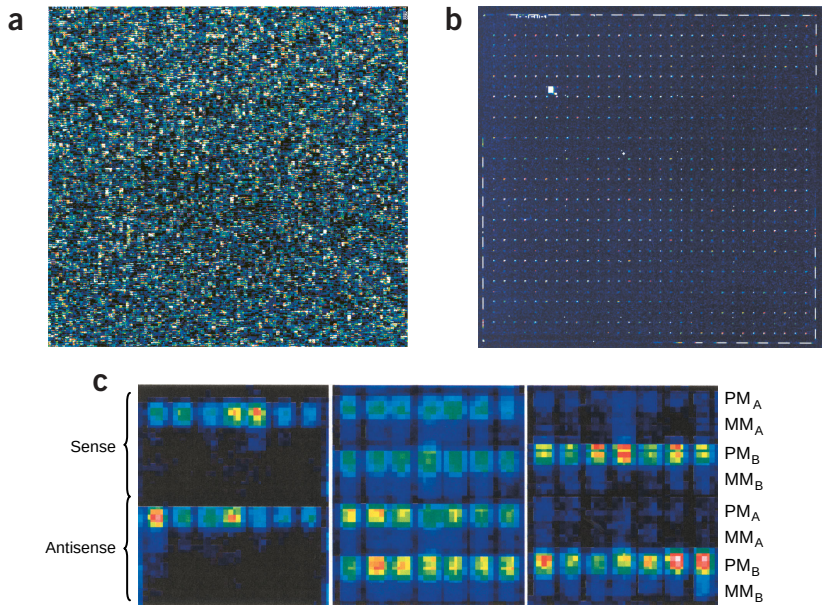

Figure 2 Hybridized chip images. (**a**) Microarray hybridized to 20 µg reduced-complexity (~4 × 10^7 bp) biotin-labeled DNA (**b**) Microarray hybridized with 20 µg biotin-labeled human genomic DNA (3 × 10^9 bp). Signals from hybridization controls are detected. (**c**) SNP miniblock showing hybridization of FSP target in three individuals, demonstrating the three possible genotypes: AA (left), AB (middle) and BB (right). Probes are synthesized as perfect-match (PM) 25-mers, and as one-base mismatches (MM) in the center. Probes for both A and B alleles on both sense and antisense strands are synthesized, for a total of 56 probes per SNP miniblock.

Consistent with theoretical predictions[21], previous results for a smaller number of SNPs in an ethnically diverse set of samples show a positive correlation between allele frequency and ancestral state[19,20]. We extended these results by examining this relationship in three human populations for a large number of SNPs. We plotted the distribution of the chimpanzee and gorilla, that is, ancestral, alleles as a function of SNP allele frequency in the African-American, Caucasian and Asian populations and found in each case a strong positive correlation; the higher the SNP allele frequency, the higher the proportion of the ancestral allele (**Fig. 4**). The slope of the line in African-Americans is 0.97, indicating a nearly one-to-one correlation between ancestral state and allele frequency, consistent with theoretical predictions[21]. In contrast, the slopes of the Caucasian and Asian populations are 0.62 and 0.52, respectively (**Fig. 4**). This indicates that in these two populations, the ancestral allele is not always the most frequent; about 20% of the time, the newer allele has become more frequent in these populations. These data are consistent with distinct demographic forces (*e.g.*, population bottlenecks and/or expansions) or recent selective events (*e.g.*, local adaptation) occurring in the Caucasian and Asian populations. This is consistent with the hypothesis that a subset of the human population migrated out of Africa and into Europe and Asia approximately 50,000 years ago[22].

This study represents a proof of principle for the simultaneous genotyping of 14,548 SNPs without locus-specific PCR or automation. With this technology, it is feasible to conduct linkage analyses, to rapidly determine allele frequencies, to discover AIMs in other populations, and to identify changes in DNA copy number in cancer. One can easily scale WGSA to genotype greater numbers of SNPs, for example, >100,000. Improvements in chip technology allow the synthesis of more SNPs per array, and improvements in target preparation allow higher complexity fractions to be genotyped. Our approach not only scales to larger numbers of SNPs, but also scales to other complex organisms. As the technology is scaled beyond 100,000 SNPs, it will be used to create haplotype maps[23] and to map regions of linkage disequilibrium across the genome, all at unprecedented resolution. With these tools, it will be possible to embark upon genetic studies aimed at uncovering the molecular basis of complex human phenotypes and to better understand the evolutionary history of our species.

METHODS

Array design. To genotype as many SNPs as possible on the smallest number of arrays, we designed the arrays to interrogate only those SNPs predicted to be amplified by our biochemical assays. Completion of the draft human genome sequence made it possible to conduct *in silico* digestions of total genomic DNA, identify fragments of the desired size and predict which SNPs should be present on those fragments. We excluded fragments containing repetitive sequences within the tiled region; these represented about 25–30% of TSC SNPs. We synthesized a series of 11 arrays containing sequences from 71,931 unique SNPs present in three different genomic subfractions (*Eco*RI, *Bgl*II and *Xba*I). A total of 56 probes were synthesized for each SNP (**Fig. 2c**). For each SNP, four probes (25-mers) were synthesized, spanning seven positions along both strands of the SNP-containing sequence, with the SNP position in the center (position zero) as well as at –4, –2, –1, +1, +3, +4. Probes were synthesized for both sense and

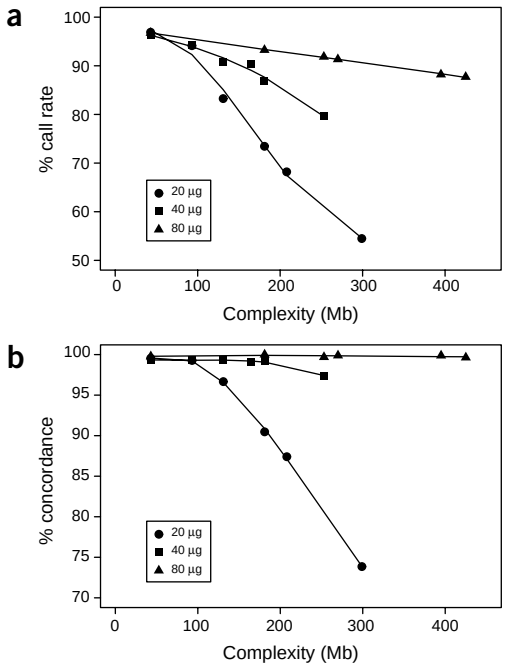

Figure 3 Effect of complexity on call rate and concordance. Genomic subfractions were prepared separately and then combined at the hybridization step to give various complexities from 43 Mb to 425 Mb. Three target DNA amounts were tested. (**a**) Genotypes were called and the call rate determined. (**b**) Percent concordancy with SBE data was determined for each point (**Supplementary Table 1**). Note the different scale of the *y*-axes in **a** and **b**.

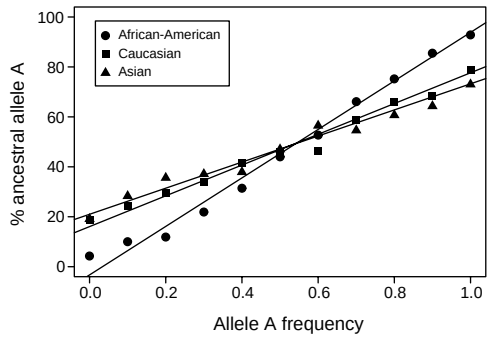

Figure 4 Percentage ancestral allele as a function of allele frequency in three populations. Genotypes were determined for chimpanzee and gorilla and the percent 'A' allele was calculated for each of 11 frequency bins. The 'A' allele for each SNP was determined alphabetically.

antisense strands. Four probes were synthesized for each of the seven positions—a perfect match (PM) for each of the two SNP alleles (A, B) and a one-base central mismatch (MM) for each of the two alleles[24] (**Fig. 2c**). These four probes are referred to as a probe quartet. It is possible to reduce the number of probes per SNP from 56 to 40 without loss of accuracy (http://www.affymetrix.com/products/arrays/specific/10k.affx). Normalized discrimination, calculated as (PM – MM)/(PM + MM), is a measure of sequence specificity and is used in the detection filter of the genotype-calling algorithm[25]. Approximately 9% of the SNPs were filtered out before classification because they had discrimination scores <0.03. This may be caused by a number of factors: *in silico* prediction errors owing to errors in the draft human genome sequence, poor amplification of loci with secondary structure, or cross-hybridization with other genomic sequences.

DNA samples. Samples used in the algorithm training set included DNA from 24 individuals from the polymorphism discovery panel (PD1-24)[26], 6 unrelated Centre Etude Polymorphism Humain (CEPH) individuals, 20 African-Americans, 20 Asians and 38 Caucasians from the TSC Allele Frequency panels, and a chimpanzee (NA03448A) and gorilla (NG05251B), all available through the Coriell Institute for Medical Research as part of the National Institute of General Medical Sciences Human Genetic Mutant Cell Repository.

Target preparation. Total genomic DNA (250 ng) was incubated with 20 units of *Eco*RI, *Bgl*II or *Xba*I restriction endonucleases (New England Biolabs (NEB)) at 37 °C for 4 h. After heat inactivation at 75 °C for 20 min, the digested DNA was incubated with 0.25 µM adaptors and DNA ligase (NEB) in standard ligation buffer (NEB) at 16 °C for 4 h. The sample was incubated at 95 °C for 5 min to inactivate the enzyme. Target amplification was performed with ligated DNA and 0.5 µM primer in PCR Buffer II (Perkin Elmer) with 2.5 mM $MgCl_2$, 250 µM dNTPs and 50 units of *Taq* polymerase (Perkin Elmer). Cycling was conducted as follows: 95 °C for 10 min, followed by 20 cycles of 95 °C for 10 s, 58 °C for 15 s, 72 °C for 15 s, followed by 25 cycles of 95 °C for 20 s, 55 °C for 15 s, 72 °C for15 s. Final extension was done at 72 °C for 7 min. The amplification products were concentrated with a YM30 column (Microcon) centrifuged at 14,000 ×*g* for 6 min. The column was washed twice with 400 µl water, respun at 14,000 ×*g* and inverted, and the sample was recovered in a clean tube by centrifuging at 3,000 ×*g* for 3 min. The sample was digested with 0.045 units DNase (Affymetrix) and 0.5 units calf intestinal phosphatase (Gibco) in RE Buffer no. 4 (NEB) at 37 °C for 30 min. Enzymes were inactivated at 95 °C for 15 min. Samples were labeled with 15–20 units terminal deoxytransferase (Promega) and 18 µM biotinylated ddATP (NEN) in TdT buffer (Promega) at 37 °C for 4 h. After heat inactivation at 95 °C for 10 min, samples were injected into microarray cartridges and hybridized overnight following manufacturer's directions (Affymetrix). Microarrays were washed in a fluidics station (Affymetrix) using 0.6 × sodium citrate phosphate EDTA Tris (SSPET), followed by a three-step staining protocol. The arrays were first incubated with 10 µg/ml streptavidin (Pierce), then washed with 6× SSPET, then incubated with 10 µg/ml biotinylated anti-streptavidin (Vector Lab) and 10 µg/ml streptavidin-phycoerythrin conjugate (Molecular Probes), and finally washed with 6× SSPET. Microarrays were scanned according to manufacturer's directions (Affymetrix). We estimate each of the enzyme fractions used in this study to have a complexity of ~43 Mb. We calculated the theoretical value for complexity based on restriction enzyme sites present in the draft human genome sequence (NCBI June 2002 release) and summed the length of all 400- to 800-bp fragments predicted to be generated after complete digestion. This estimate of complexity is affected by several factors: accuracy of genome sequence used for *in silico* fractionations, completion of restriction enzyme digestion, efficiency of adaptor ligation and uniformity of amplification of 400- to 800-bp size fragments. For complexity titration experiments we used the following estimated complexity fractions in various combinations to give the final complexities shown on the *x*-axes of **Fig. 3**: *Xba*I (43 Mb), *Eco*RI (43 Mb), *Bgl*II (41 Mb), *Nco*I (41 Mb) *Bsr*GI (47 Mb), *Hin*dIII (50 Mb), *Nde*I (59 Mb), *Hin*P1 (122 Mb), *Hpa*II (210 Mb). Fractions were amplified, fragmented and labeled separately, then combined in equimolar amounts before the hybridization step.

Algorithm training. We developed an automated scoring process for calling genotypes[11]. The training data were derived from 108 DNA samples derived from ethnically diverse individuals. We calculated relative allele signal (RAS) values for each SNP on both the sense and antisense strands and plotted them for all 108 individuals in two dimensions[11]. Some SNPs showed three clearly defined clusters (**Supplementary Fig. 1a**), whereas others showed more diffuse clusters (**Supplementary Fig. 1b**) or no clear clusters at all (**Supplementary Fig. 1c**). For those SNPs having lower minor allele frequencies, the genotypes fell into only two clusters, with the minor allele homozygote clusters being absent (**Supplementary Fig. 1d**). After graphical visualization of clusters derived from RAS values in two dimensions, we developed an algorithm using a modification of partitioning around medoids[11], to classify these points into two or three clusters. We evaluated the quality of classification by calculating the average silhouette width, *s*[27]. As *s* approaches 1.0, clusters are tight and well separated, whereas low values of *s*—for example, <0.5—are derived from poorly clustering SNPs.

We developed a series of conservative heuristics for ranking the SNPs according to their clustering properties. Of the 71,931 SNPs assessed in this experiment across three enzyme fractions, 14,548 formed three clusters with $s > 0.7$ and showed separation of RAS medians between clusters >0.2, and 90% of the samples passed the detection filter. We scored only SNPs with high silhouette scores that formed three clusters containing at least two points in each cluster; therefore, many SNPs that formed three clusters with only one instance of the minor allele homozygote, or SNPs that formed only two good clusters (**Supplementary Fig. 1d**), did not meet the cutoff criteria. Respectively, 5% and 21% of the SNPs fell into these two latter categories. Genotyping 108 individuals limits the ascertainment of SNPs to those with a minor allele frequency of >10% and results in a bias towards SNPs with high heterozygosities. A further bias was introduced because we did not constrain all three clusters for a given SNP to come from individuals in the same ethnic group; thus, all three genotypes are not necessarily observed in any one ethnic group.

Reproducibility and accuracy. Reproducibility was determined on a set of 38 samples, genotyped as incoming data on clusters defined by the training set. The percentage of successful genotype calls (call rate) was averaged over 38 samples and ranged from 91.1% to 97.9% (**Supplementary Table 1**). The average genotype call rate was 95.8% ± 1.2% (mean ± s.d.), demonstrating a high level of reproducibility (**Supplementary Table 1**). We also studied SNPs present in multiple enzyme fractions and tiled on two or more arrays. Of the 205 SNPs synthesized on two or more arrays and captured by different enzyme fractions, the concordance rate for genotype calls was 99.5% across 30 individuals (data not shown).

We determined the accuracy of our genotype calls in two ways: directly, through the use of genotypes obtained by other technologies, and indirectly, by genotyping complete hydatidiform moles and calculating mendelian inheritance errors in families. Hydatidiform moles are products of abnormal conception arising from the fertilization of an empty ovum by a single sperm, resulting in complete duplication of the haploid paternal genome. We first obtained reference genotypes for approximately 1,000 SNPs assayed using single-base extension (SBE) technology[28] and compared these genotypes to those gener-

ated by WGSA (**Supplementary Table 1**). We found a concordance rate of 99.15% in these markers over 38 samples (total of 39,848 calls compared). Ten SNPs accounted for >50% of the 339 discordant genotypes. We obtained *de novo* nucleotide sequences for these ten SNPs across individuals exhibiting discordant genotypes, and found that WGSA genotype calls were concordant with sequence data 44% of the time. Thus, the accuracy of WGSA genotype calls is most likely >99.5%. We also compared genotypes for 65 SNPs across seven individuals with data derived from high-resolution scanning of chromosome 21 (ref. 29). Of 287 calls compared between the two datasets, there was only one discordant genotype (that is, concordance rate = 99.6%). Additional confidence in the accuracy of our genotype calls was obtained indirectly by examining genomic DNA isolated from two complete hydatidiform moles. Genotypes are expected to be homozygous for all markers[30]. Both tumors showed 0.4% heterozygosity, consistent with expectations of a completely duplicated haploid genome, whereas a control sample of normal placenta showed 35.3% heterozygosity (data not shown). Lastly, we determined the mendelian inheritance error rate by genotyping samples from CEPH families for a subset of the markers (7,005 SNPs) and determined it to be <0.3% (unpublished data). Accuracy and reproducibility can be further increased by removing SNPs that genotype across fewer than 85% of samples and that show mendelian errors in more than one family (http://www.affymetrix.com/products/arrays/specific/10k.affx).

Allele frequency determinations. Of the 14,548 SNPs scored in each of the three populations, we included only those SNPs that scored in at least 75% of individuals in each population (13,647 SNPs). A comparison of allele frequencies derived from a set of 20 Caucasians versus a set of 38 Caucasians shows a high correlation ($R^2 = 0.96$) between them, indicating that sampling of 40 chromosomes provides stable estimates of allele frequencies for these SNPs in that population.

Marker distribution. The SNPs were mapped on the human genome sequence by TSC. The distribution of inter-SNP distances is shown in **Supplementary Figure 3** online; the mean and median values are 174 kb and 80.8 kb, respectively. Of these markers, 5,058 are spaced at distances of 50 kb or less; 3,868 are spaced at distances of 25 kb or less. The chromosomal distribution of SNPs is shown in **Supplementary Figure 4** online. As expected, regions of the genome known to have fewer SNPs, for example, telomeres, centromeres and regions of heterochromatin, also contain fewer WGSA SNPs.

Availability of data. Genotype data for 14,548 SNPs on 108 individuals (including 60 individuals from the allele frequency study), as well as chimpanzee and gorilla genotypes, have been deposited in dbSNP at http://ncbi.nlm.nih.gov/SNP/ under the handle "AFFY."

URLs. The SNP consortium website that describes the allele frequency samples is: http://snp.cshl.org/allele_frequency_project/panels.shtml.

The Coriell Institute for Medical Research as part of the National Institute of General Medical Sciences Human Genetic Mutant Cell Repository website is http://locus.umdnj.edu/nigms/.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We thank David Altshuler, Eric Lander, Thomas Gingeras, Richard Rava, Michael Shapero and Jon McAuliffe for helpful suggestions and critical reading of the manuscript.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests (see the *Nature Biotechnology* website for details).

Received 7 January 2003; accepted 16 July 2003
Published online at http://www.nature.com/naturebiotechnology/

1. Ardlie, K.G., Kruglyak, L. & Seielstad, M. Patterns of linkage disequilibrium in the human genome. *Nat. Rev. Genet.* **3**, 299–309 (2002).
2. Sachidanandam, R. *et al.* The International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
3. Kwok, P.-Y. Methods for genotyping single nucleotide polymorphisms. *Annu. Rev. Genomics Hum. Genet.* **2**, 235–258 (2001).
4. Syvanen, A.-C. Accessing genetic variation: genotyping single nucleotide polymorphisms *Nat. Rev. Genet.* **2**, 930–942 (2001).
5. Lipshutz, R.J., Fodor, S.P., Gingeras, T.R. & Lockhart, D.J. High density synthetic oligonucleotide arrays. *Nat. Genet.* **21**(1 Suppl), 20–24 (1999).
6. Lisitsyn, N., Lisitsyn, N. & Wigler, M. Cloning the differences between two complex genomes. *Science* **259**, 946–51 (1993).
7. Lucito, R. *et al.* Genetic analysis using genomic representations. *Proc. Natl. Acad. Sci. USA* **95**, 4487–4492 (1998).
8. Vos, P. *et al.* AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res.* **23**, 4407–4414 (1995).
9. Altshuler, D. *et al.* An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**, 513–516 (2000).
10. Dong, S. *et al.* Flexible use of high-density oligonucleotide arrays for single-nucleotide polymorphism discovery and validation. *Genome Res.* **11**, 1418–1424 (2001).
11. Liu, W.-m. *et al.* Algorithms for large scale genotyping microarrays. *Bioinformatics*, (2003), in the press.
12. Weir, B.S. *Genetic Data Analysis II* (Sinauer Associates, Sunderland, Massachusetts, 1996).
13. Bowcock, A.M. *et al.* Drift, admixture, and selection in human evolution: a study with DNA polymorphisms. *Proc. Nat. Acad. Sci. USA* **88**, 839–843 (1991).
14. Collins-Schramm, H. *et al.* Ethnic-difference markers for use in mapping by admixture linkage disequilibrium. *Am. J. Hum. Genet.* **70**, 737–750 (2002).
15. Briscoe, D., Stephens, J.C. & O'Brien, S.J. Linkage disequilibrium in admixed populations: applications in gene mapping. *J. Hered.* **85**, 59–63 (1994).
16. Parra, E.J. *et al.* Estimating African-American admixture proportions by use of population-specific alleles. *Am. J. Hum. Genet.* **63**, 1839–1851 (1998).
17. McKeigue, P.M., Carpenter, J.R., Parra, E.J. & Shriver, M.D. Estimation of admixture and detection of linkage in admixed populations by a Bayesian approach: application to African-American populations. *Ann. Hum. Genet.* **64**, 171–186 (2000).
18. Hacia, J.G. Genome of the apes. *Trends Genet.* **17**, 637–645 (2001).
19. Hacia, J.G. *et al.* Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. *Nat. Genet.* **22**, 164–167 (1999).
20. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* **22**, 231–238 (1999).
21. Watterson, G.A. & Guess, H.A. Is the most frequent allele the oldest? *Theor. Pop. Biol.* **11**, 141–160 (1977).
22. Cavalli-Sforza, L.L., Menozzi, P. & Piazza, A. *The History and Geography of Human Genes* (Princeton University Press, Princeton, NJ, 1994).
23. Gabriel, S.B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).
24. Lindblad-Toh, K. *et al.* Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays. *Nat. Biotechnol.* **18**, 1001–1005 (2000).
25. Liu, W.-m. et al. Rank-based algorithms for analysis of microarrays. in Microarrays: Optical Technologies and Informatics (eds. Bittner, M.L., Chen, Y., Dorsel, A.N. & Dougherty, E.R.) Proc. SPIE **4266**, 56–67 (2001).
26. Collins, F.S., Brooks, L.D. & Chakravarti, A. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.* **8**, 1229–1231 (1998).
27. Rousseeuw, P.J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
28. Picoult-Newberg, L. *et al.* Mining SNPs from EST databases. *Genome Res.* **9**, 167–174 (1999).
29. Patil, N. *et al.* Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719–1723 (2001).
30. Fan, J.-B. *et al.* Paternal origins of complete hydatidiform moles proven by whole genome single-nucleotide phaplotyping. *Genomics* **79**, 58–62 (2002).